Journal of Nuclear Medicine, published on June 30, 2022 as doi:10.2967/jnumed.122.264063

Fully-automated, semantic segmentation of whole-body ^{18}F-FDG PET/CT images based on data-centric artificial intelligence

Lalith Kumar Shiyam Sundar[1*], Josef Yu[1,2*], Otto Muzik[3], Oana C. Kulterer[2], Barbara Fueger[2], Daria Kifjak[2,4], Thomas Nakuz[2], Hyung Min Shin[5], Annika Katharina Sima[2], Daniela Kitzmantl[2], Ramsey D Badawi[6], Lorenzo Nardo[6], Simon R Cherry[6], Benjamin A Spencer[6], Marcus Hacker[2] and Thomas Beyer[1]

[1]Quantitative Imaging and Medical Physics (QIMP) Team, Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Austria

[2]Department of Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Vienna, Austria

[3]Department of Pediatrics, Wayne State University School of Medicine, Children's Hospital of Michigan, Detroit, MI, USA

[4]Department of Radiology, University of Massachusetts Chan Medical School/UMass Memorial Health Care, Worcester, MA, USA

[5]Division of General Surgery, Department of Surgery, Medical University of Vienna, Vienna, Austria

[6]Department of Biomedical Engineering and Radiology, University of California-Davis, Davis, CA, USA

Corresponding Author

Otto Muzik, PhD

Wayne State University School of Medicine, Detroit, Michigan, USA

Telephone: +1-(734)-223-3506

otto@pet.wayne.edu

First authors contact information:

Lalith Kumar Shiyam Sundar, PhD and Josef Yu MD (*Equal contribution)

QIMP Team, Center for Medical Physics and Biomedical Engineering, Medical University of Vienna,

Vienna, Austria

Telephone: +43 1 40400 55450

Lalith.shiyamsundar@meduniwien.ac.at

Running title: PET/CT Multi-organ segmentation

Word count: 4624

Figures: 6

Keywords: multi-organ segmentation, total-body PET, systems medicine, artificial neural networks, automated segmentation

Funding

This work was supported in parts by NIH Research grant R01CA29422.

ABSTRACT

We introduce Multi-Organ Objective Segmentation (MOOSE) software that generates subject-specific, multi-organ segmentation using data-centric AI principles to facilitate high-throughput systemic investigations of the human body via whole-body PET imaging. **Methods:** Image data from two PET/CT systems (uEXPLORER and Siemens TruePoint TrueView) was used in training MOOSE. For non-cerebral structures, 50 WB-CT images were used, 30 of which were acquired from healthy controls (HC, 14M/16F) and 20 datasets were acquired from oncology patients (14M/6F). Non-cerebral tissues consisted of 13 abdominal organs, 20 bone segments, subcutaneous fat, visceral fat, psoas, and skeletal muscle. An expert panel performed manual segmentation of all non-cerebral structures except for subcutaneous, visceral fat, and skeletal muscle, which were semi-automatically segmented using thresholding. A majority-voting algorithm was used to generate a 'reference standard' segmentation. From the 50 CT datasets, 40 were used for training and 10 for testing purposes. For cerebral structures, 34 ^{18}F-FDG PET/MRI brain image volumes were used from 10 HC (5M/5F imaged twice) and 14 non-lesional epilepsy patients (7M/7F). Only ^{18}F-FDG PET images were considered for training: 24/34 and 10/34 volumes were used for training and testing. The dice score coefficient (DSC) was used as the primary and the average symmetric surface distance (ASSD) as a secondary metric to evaluate the automated segmentation performance. **Results:** An excellent overlap between the reference labels and MOOSE-derived organ segmentations was observed: 92% of non-cerebral tissues showed DSC values of >0.90, while a few organs exhibited lower DSC values (e.g., adrenal glands (0.72), pancreas (0.85), and bladder (0.86)). The median DSC values of brain subregions derived from PET images were lower. Only 29% of the brain segments had a median DSC of >0.90, while segmentation of 60% of regions yielded a median DSC of 0.80-0.89. Results of the ASSD analysis demonstrated that the average distance between the reference standard and the automatically segmented tissue surfaces (organs, bones, brain regions) lies within the size of image voxels (2mm).

Conclusion: The proposed segmentation pipeline allows automatic segmentation of 120 unique tissues from whole-body ^{18}F-FDG PET/CT images with high accuracy.

INTRODUCTION

Living organisms maintain steady internal physiological conditions through dynamic, self-regulating multi-organ systemic interactions (*1*), also known as *homeostasis*. In healthy subjects, any notable deviation from homeostasis is avoided with the aid of systemic feedback loops (*2*). Chronic pathologies are conceived as sustained disturbances in homeostasis, which cannot be compensated by systemic communications (*3*). Molecular imaging modalities, such as positron emission tomography (PET), can provide essential insights into diverse biological processes within the human body by using highly-specific radiotracers that track molecular function *in vivo* (*4*). Assuming that homeostasis is associated with a balanced, albeit variable, glycolytic pattern, PET can help characterise bespoke feedback loops and deviations that lead to pathologies. However, until recently, whole-body PET imaging protocols were typically limited to only a portion of the patient's body (e.g., neck to upper thigh) due to the relatively narrow axial field-of-view (FOV, 15-25cm) of PET systems. This limitation required multiple bed positions to be acquired sequentially to cover the axial field of investigation. Nonetheless, this acquisition mode failed to fully harness the multi-systemic physiological information provided by PET imaging (*5*).

With the recent advent of large axial FOV PET/CT systems (>70 cm) (*6*–*8*), the opportunity arose to acquire total-body (TB) PET images with only 1-2 bed positions, facilitating multi-organ system analysis. Such systemic analysis might allow the investigation of multi-organ interactions in various pathologies, such as those associated with cancer (*9*), cachexia (*10*,*11*), metabolic syndrome (*12*), or the more recent COVID-19 virus (*13*). However, the amount of data generated by this new generation of PET/CT systems are too large to be adequately analysed without automated processing pipelines.

In response, we developed a multi-organ objective segmentation (MOOSE) tool, an open-source software framework based on data-centric Artificial Intelligence (*14*) (AI) principles (Supplemental Fig. 1) to allow fully-automated generation of a subject-specific total-body (TB) ^{18}F-FDG PET/CT tissue-map consisting of over 100 different tissue types. The development of such a software tool dramatically

increases the amount of information that can be efficiently extracted from PET data. Further, such a tool provides a means to observe normal physiology and pathological conditions globally, permitting systems-level investigations into human physiology. For example, when applied in a clinical setting, this approach will allow physicians to automatically generate a list of standard uptake values (SUV) for all organs of interest, which might provide auxiliary information during the diagnostic process. In addition, the automated generation of a complete set of organ-specific SUVs lends itself well to AI-supported diagnostic screening, allowing organ SUV ratios to be compared across subjects and alerting the physician about potential secondary pathologies.

MATERIALS AND METHODS

Overall segmentation strategy

Our approach is based on the latest state-of-the-art nnU-Net segmentation framework (*15*). More importantly, we propose a data-centric approach (*14*,*16*) where the network model is fixed, and the data is iteratively augmented to increase the performance of the AI system. As such, the model's performance is continuously monitored. As new data deviating from the training dataset's characteristics enter the processing stream, the model is retrained to enhance performance.

Data

Two different types of datasets were used for the development of a software tool able to segment both cerebral (83 regions) and non-cerebral structures (37 tissues).

For training and evaluation of non-cerebral structures, 50 whole-body low-dose CT datasets were used. Among these 50 datasets, 30 CT images were acquired from healthy volunteers (14M/16F, 47 ± 13 years) using the uEXPLORER total-body PET/CT system (*17*). The remaining 20 datasets belonged to a retrospective patient cohort from a Siemens TruePoint TrueView (TPTV) PET/CT system (14M/6F, 67 ± 12 years). The non-cerebral tissues atlas consists of 13 abdominal organs, 20 bone segments, subcutaneous fat, visceral fat, psoas, and skeletal muscle (Supplemental Table 1, Supplemental Fig. 2).

An expert segmentation panel comprised of four physicians and four medical students (final year) was responsible for the manual segmentation of all non-cerebral structures, except for subcutaneous and visceral fat and skeletal muscle, which were outlined using an established thresholding method (*18*). The physicians were responsible for segmenting the abdominal organs and psoas while the students generated the bone segments. From the 50 datasets, 40 were used for training, and 1tenwere used for testing (hold-out dataset) purposes.

For training and evaluation of cerebral structures, we used 34 ^{18}F-FDG PET/MRI brain datasets (10 healthy controls test-retest: 5M/5F, 27 ± 7 years and 14 non-lesional epilepsy patients: 7M/7F, 29 ± 9 years)

(*19*,*20*). The cerebral atlas consisted of 83 brain subregions (Supplemental Table 1), automatically created from PET data in combination with T1-MR images and the Hammersmith atlas (*21*). In short, subject-specific T1-MR images were normalized to MNI space using SPM 12 (*22*). The obtained (inverse) transform was then used to spatially transform brain regions of the Hammersmith atlas into the individual subject's native space, yielding 83 subject-specific cerebral sub-regions which were transfered to coregistered PET image volumes. Of the 34 datasets, 24 and 10 were used for training and testing, respectively.

Reference standard generation

To address inter-variability issues of the organ segmentation, the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (*23*) was employed to generate reference volumes for further performance assessment. Each reference volume represents a probabilistic estimate of the “true” segmentation as well as a measure of multi-operator segmentation performance, (STAPLE variance). The STAPLE method was not employed for reference segmentations derived using automatic (brain atlas) or semi-automatic (thresholding) methods.

U-Net-based semantic segmentation

The nnU-Net implementation of the generic U-Net architecture is a self-configuring method for deep learning-based biomedical image segmentation. This implementation exhibits strong performance by retaining the original U-Net-like architecture while automating the complex process of manual hyperparameter configuration (*15*).

In our implementation, training of the nnU-Net was performed separately for the following four structure classes: (i) 13 abdominal organs and psoas, (ii) 20 bone structures, (iii) 83 brain regions (iv) fat (subcutaneous and visceral) and skeletal muscle. Segmentation of non-cerebral tissues was performed based on CT data, whereas segmentation of cerebral regions was carried out using ^{18}F-FDG PET images.

Assessment of deviation from training dataset distribution

It is unlikely that any training dataset will be sufficient to fully capture the variability encountered in clinical routine. Accordingly, a data-centric approach is necessary, permitting continuous monitoring of segmentation performance so that data that substantially deviates from the original training data distribution (i.e., Out-of-distribution (OOD) data) is detected. Erroneous segmentation results obtained for such data will then require manual correction by a human expert. Once corrected, this data can be appended in suitable quantities to the initial training dataset for retraining purposes.

Since continuous operator-based monitoring of segmentation performance is untenable in clinical routine, we developed an automated error analysis routine that detects OOD datasets based on morphometric analysis of organ shapes (e.g., elongation, volume, area, maximum and minimum bounding box diameter), which were determined for each STAPLE-derived segmentation of structures, and a normative morphological feature database was generated. Upon segmentation of a new dataset, morphological features for each segmented structure are calculated and compared to the normative morphology database, yielding a distance (Z-score) in "similarity space" for each structure. The Z-score reflects the difference between the shapes of the segmented structure in comparison to its normative value obtained from the training datasets. In our implementation, we chose a value of Z = 1.5 as the cut-off for OOD labelling.

Algorithm performance vs. training sample size

Primary performance assessment of the automated segmentation (MOOSE) was performed for all structures using the Dice score coefficient (DSC) (*24*). A DSC value of 1.0 with respect to STAPLE indicates perfect overlap and 0 indicates no overlap. In addition, the average symmetric surface distance (ASSD) (*25*) was used as a secondary metric, representing the average distance (in mm) between surface voxels of the standard and the automated segmentation.

To assess the segmentation performance as a function of training sample size, we calculated for each non-cerebral structure the DSC and the ASSD using the segmented volumes derived using 10 (D10),

20 (D20), and 40 (D40) training data sets, respectively. A similar analysis was performed for cerebral regions with 8 (D'8), 16 (D'16), and 24 (D'24) data sets. In both instances, cases were randomly selected from the whole datasets (50 cases for non-cerebral and 34 cases for cerebral structures). The testing (hold-out) dataset included 10 cases that were not part of the training sets in both instances.

Algorithm performance vs. training dataset variability

To investigate the effect of training dataset variability on segmentation performance, we performed a series of training/test runs using various mixtures of two datasets that differed significantly with respect to arm position (either arms down or crossed on chest, Supplemental Fig. 3). We created four subsets of training datasets, each with a total sample size of 20. The first dataset consisted of 20 low-dose CT images with arms down (SMS20). The other three training datasets included mixtures of images: (i) MIX2-18 (18 arms down, 2 crossed), MIX5-15 (15 arms down, 5 crossed), and MIX10-10 (10 arms down, 10 crossed). Networks trained on these four training datasets were then used to segment 10 test datasets that included only images with crossed arm positions (X10). The following four (training test) scenarios were investigated: [SMS20→X10], [MIX2-18→X10], [MIX5-15→X10] and [MIX10-10→X10]. Segmentation results were assessed separately for bone structures of the arm (radius, ulna, carpal, metacarpal, fingers) and for all other bone structures (that did not differ positionally). This analysis provided information regarding the necessary variability in the training dataset required to segment OOD data accurately.

Algorithm performance for clinical OOD datasets

We applied the trained network to two small pathological cohorts that were not part of the initial training set: three lymphoma and three mesothelioma lung cancer cases. The intent was to assess the performance of MOOSE on clinical datasets that differ significantly from the training data distribution. The segmentation quality was evaluated based on similarity space analysis (Z-scores). OOD datasets with incorrect segmentations were manually corrected, and the corrected segmentations were then appended to the original training datasets for retraining purposes.

Statistical assessment

A paired t-test was applied to determine whether DSC values differed significantly between the various training sample sizes and to investigate the effect of training dataset variability (either fully OOD or mixed) on DSC values. In addition, a correlation analysis (Pearson's rho) was performed to investigate the relationship between STAPLE variance and the DSC values associated with the best (D40) training sample size. A similar analysis was also performed using the ASSD metric.

Software tool implementation

Our processing pipeline is based on Python, and C++ programming languages, with the nnU-Net framework representing the segmentation backbone, built using PyTorch 1.6.031 (*26*). Similarity space was implemented using the morphometric capabilities of SimpleITK 2.1.0 (*27*), and manual cleaning of erroneous segmentation results was performed using 3D Slicer (*28*) (Version 4.11.20210226).

RESULTS

Effect of training data size on segmentation performance

The majority of non-cerebral tissues (81%) were segmented with high accuracy (DSC>0.90) as seen from Fig. 1**.** Of note, DSC analysis generally showed an excellent overlap between STAPLE-derived reference and organ segmentations based on 10 (D10), 20 (D20), and 40 (D40) training data sets. This excellent overlap was confirmed based on ASSD analysis, yielding average distances of 1.40 $\pm$ 1.29, 1.05 $\pm$ 1.26 and 0.68 $\pm$ 0.52mm for D10, D20 and D40, respectively. However, the performance of the automated segmentation was suboptimal for a small group of organs (highlighted in red, Fig. 1), with low median DSC and high SD values (Supplemental Fig. 4A), such as the adrenal glands (DSC=0.72), pancreas (0.85), and bladder (0.86). Subsequent correlation analysis of the STAPLE variance and the DSC values derived from the (best) D40 training set is shown in Supplemental Fig. 4B. The graph indicates an overall highly significant negative correlation (rho = -0.79, p = 0.002) with the three identified regions showing high STAPLE variance. This significant correlation with the STAPLE variance was also reproduced using the ASSD metric (rho = 0.60, p = 0.042, Supplemental Fig. 5), indicating that accurate segmentation of this subset of regions is challenging even for human experts.

The segmentation performance for bone structures was similar to that of the abdominal organs (Fig. 2). Again, one notes an excellent overlap between the reference structure volumes and those obtained using the automated segmentation based on 10 (D10), 20 (D20), and 40 (D40) training data sets (ASSD of 1.63 $\pm$ 3.01, 1.61 $\pm$ 3.14 and 0.83 $\pm$ 0.76mm, respectively), except for a small number of bone structures with either low DSC mean or high SD values (Supplemental Fig. 6). These structures were the carpal bones, metacarpal bones, and phalanges of the toes. Removal of these organs resulted in a similar segmentation performance between D20 and D40 (p = 0.07), with segmentation based on D10 remaining significantly worse than D20 (p = 0.016) and D40 (p = 0.010).

Although median DSC values of brain subregions derived from PET images were relatively low (only 29% of brain segments had median DSC values >0.90, see Supplemental Fig. 7), ASSD values showed sub-voxel differences between the template regions and the automated segmentation, with similar performance across the D10 (0.52 $\pm$ 0.35mm), D20 (0.53 $\pm$ 0.41mm) and D40 (0.46 $\pm$ 0.27mm) datasets.

Effect of training dataset variability on segmentation performance

Results of dataset variability analysis are shown in Fig. 3. The figure indicates that segmentation of structures that substantially deviate from the 'expected' position in the training datasets was suboptimal. However, by including at least two cases that match the deviant position to the training dataset resulted in a significant performance improvement. Specifically, DSC values for bones of the arm were significantly lower for the fully OOD [SMS20→X10] scenario (DSC = 0.87 $\pm$ 0.12) as compared to the three scenarios that included 10% [MIX2-18→X10] (DSC = 0.92 $\pm$ 0.06; p = 0.04), 25% [MIX5-15→X10] (DSC = 0.940 + 0.003; p = 0.01) and 50% [MIX10-10→X10] (DSC = 0.91 + 0.04; p = 0.04) of cases that matched the deviant position. In addition, the COV for DSC values derived from the three mixed training datasets was significantly lower (COV = 6.6%, p = 0.01; COV = 3.3%, p = 0.03; COV = 4.3%. p = 0.01) as compared to the COV for DSC values derived using the fully OOD training dataset (COV = 13.5%). In comparison, performance of all four scenarios for bone structures that were matched in position between the training and test datasets was similar, with DSC values of >0.95 (Fig. 3).

Detection of OOD segmentation errors

Application of similarity space analysis identified segmentation errors in clinical datasets that included various anatomical pathologies, representing OOD datasets for specific organs. This was clearly demonstrated by applying the initially trained neural network to two distinct OOD datasets (lymphoma and mesothelioma) that were not part of the initial training set. Specifically, all lymphoma patients presented with splenomegaly which led to its incorrect classification as a "liver" and "spleen" (Fig. 4A). Following manual correction (time required ~3min per case), we appended two corrected data sets to the

252 original training set to retrain the neural network. The retrained neural network correctly segmented the
253 abnormally enlarged spleen in the third lymphoma patient which was used as a hold-out dataset (Fig. 5).
254 Similarly, the large tumor mass in the lungs of mesothelioma patients was incorrectly classified as
255 part of the liver, heart, and bladder (Fig. 4B). Again, similarity space analysis identified the incorrect
256 segmentation and labelled the dataset as representing an OOD image pattern (Fig. 6A). Following manual
257 correction of two out of three patients, these two cases were again appended to the training data set,
258 and the neural network was retrained using the extended training set. Once again, we determined an
259 improvement in the segmentation performance of the third (uncorrected) data set (Fig. 6).
260

DISCUSSION

Hybrid molecular imaging modalities, such as ^{18}F-FDG PET/CT allow the investigation of multi-organ systemic interactions through which living organisms maintain homeostasis and allostasis. The resulting images are not mere pictures - they represent a rich pallet of multi-dimensional data (*29*). By systemically parcellating these datasets into respective organ/tissue classes, one can, in theory, study system-level interactions in detail between the various homeostatic and allostatic networks, allowing a better understanding of pathological abnormalities in vivo. Nevertheless, manual segmentation of various tissues in the human body is not tenable, either in research applications or in clinical routine.

To bridge this gap, we developed a fully-automated segmentation pipeline (MOOSE) that allows the creation of subject-specific multi-tissue FDG PET/CT atlases (Supplemental Fig. 2). These tissue-maps enables the extraction of subject-specific functional information from molecular imaging data with minimal additional effort for further analysis. We based the developed segmentation pipeline on the latest state-of-art nnU-Net architecture (*15*) and demonstrated that robust training of the convolutional neural network could be achieved with as low as 20 datasets, provided that sufficient variability in the training dataset is present. In addition, our results support the concept of data-centric AI, which focuses primarily on data quality rather than quantity.

In general, MOOSE performed reasonably well in segmenting most of the non-cerebral tissues while exhibiting poorer segmentation performance on selected organs, such as thyroid, adrenal gland and bladder. Our correlation analysis revealed a significant negative correlation between the STAPLE variance and the DSC values derived from the (best) D40 training set (Supplemental Fig. 4B). This result suggests that, due to a combination of small organ size, low contrast and increased noise levels present in low-dose CT images, accurate segmentation of bespoke structures is challenging even for human experts.

AI, PET imaging, and systems biology

The ultimate objective of the developed multi-organ and tissue segmentation methodology is to promote the concept of whole-person research (*30*) and systems biomedicine (*31*) through whole/total-body ^{18}F-FDG PET/CT imaging. With the advent of large axial FOV PET/CT systems, most or all organs can be simultaneously imaged, therefore allowing an improved evaluation of interactions between organs in both healthy and diseased states. We envision that through automated extraction of rich physiological information inherent in PET/CT data (e.g., organ SUVs), disease-specific "metabolic fingerprints" can be derived that uniquely characterize diverse pathologies affecting system-level organ interaction (Supplemental Fig. 8). Such an analysis might uncover metabolic dependencies among sets of organs and might provide novel insights into metabolic pathway dysregulation associated with disease progression. Moreover, given the fact that non-cerebral tissues are segmented directly from CT data, this technique is insensitive to variations in PET tracer uptake patterns, thus allowing the study of diverse system-level functional processes using a multitude of function-specific radiotracers.

Training of neural networks using sparse datasets

It is commonly assumed that the performance of a neural network increases with the size of the training set. Therefore, most non-healthcare image classification applications are trained on vast numbers of training cases (e.g., ImageNet (*32*)). However, creating large training datasets in the medical field is problematic, as manual curation of medical images is highly time-consuming and heavily dependent on domain-specific human expert knowledge. In this study, 50 medical image datasets were manually segmented (into 120 objects for each dataset) by medical professionals. This process required significant personal effort by each expert and took several months to complete. Such an effort cannot be expected to be repeated numerous times when additional silos of data (possibly with a different distribution) become available.

In recognition of this methodological constraint, we investigated the effect of both training sample size and training sample variability on segmentation performance. Our results demonstrate that

segmentation performance is primarily dependent on whether the training dataset allows the correct identification of several distinct unique image patterns, each characterized by a mean spatial pattern and the associated variance (**Fig 4-6**). This insight also explains why more cases are usually preferred, as it is likely that a greater number of unique image patterns can be captured using a larger dataset. However, the number of images needed per unique pattern is not evident. Our results suggest that accurate segmentation of abnormal image patterns is viable, provided that the training data includes a small number (2-4) of cases that establish a distinct image pattern with the associated morphological variance.

A data-centric approach to segmentation

Over the long run, any clinically viable medical image segmentation method will require a system where incoming data is constantly utilized to adjust model parameters to accommodate changing data distributions. To accomplish this, the implemented data-centric approach executes two operations: first, it performs active monitoring of segmentation performance followed by the users input to correct the erroneous segmentation, and second, implements periodic updating of model parameters through retraining of the neural network using an updated training set (which includes the manually corrected OOD data).

In particular, segmentation performance is continuously monitored in similarity space, and feedback regarding segmentation accuracy is provided to the physician in the form of tissue-specific Z-scores that signal potential deviations from tissue shape/position in the normative training data distribution. Based on this analysis, all tissues that are judged to be out-of-distribution ($Z > 1.5$) are flagged, and the physician is prompted for corrective action. This approach ensures adequate segmentation of all tissues present in abnormal datasets and provides important curated data for future retraining of the neural network. Moreover, this strategy addresses potential segmentation problems right when they occur in the processing pipeline, when corrective actions can be carried out most efficiently and with the least effort.

It is important to note that the presented segmentation framework bears its challenges. First and foremost, this methodology mandates a high-performance workstation which might be cost-prohibitive. Our network training was performed on a dedicated server (Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz 32 CPU cores, 256 GB of RAM, and 1 x Nvidia GeForce RTX 3090 Ti), allowing the generation of one TB ^{18}FDG PET/CT tissue-map from an individual whole-body PET/CT dataset in about 30 minutes. Moreover, once OOD data sets are collected, the neural network needs to be retrained, which took approximately two days to complete using the above server configuration. Finally, there is some unavoidable subjectivity in identifying OOD data sets as the cut-off defining OOD data is based on heuristics.

CONCLUSION

We present here a fully-automated, data-centric segmentation pipeline for the creation of a total-body ^{18}F-FDG PET/CT tissue-map. The generated map is modular and consists of 120 tissues and bone structures, enabling the automated extraction of image information for both cerebral and non-cerebral regions, potentially providing added information about secondary abnormalities during the diagnostic process.

CODE-AVAILABILITY

We have named our software pipeline MOOSE[120], which stands for Multi-Organ Objective Segmentation. MOOSE[120] is free and is an open-source software. All codes related to MOOSE[120] are available online. All models for our application are publicly available, and a complete description of the processing pipeline is available on our GitHub page (https://github.com/QIMP-Team/MOOSE).

CONFLICT OF INTEREST

No potential conflicts of interest relevant to this article exist.

ACKNOWLEDGMENTS

We thank Jakub Gawrylkowicz and Sebastian Gutschmayer for their support in packaging the software. We extend our gratitude to Clemens Spielvogel for helpful discussions throughout the project. We thank Kilian Kluge for organising the mesothelioma datasets, and finally, we thank Irene Buvat for providing the lymphoma datasets to verify our hypothesis. We extend our gratitude to IBM, as this research is supported through the IBM University Cloud Award.

KEY POINTS:

QUESTION: How can we efficiently extract diagnostic information from whole-body ^{18}F-FDG-PET/CT data?

PERTINENT FINDINGS: An automated approach for multi-organ segmentation of whole-body 18F-FDG-PET data is presented. It builds on the nnU-Net methodology driven by data-centric principles and supports accurate segmentation of 37 extra-cerebral and 83 cerebral regions. Over 92% of the non-cerebral tissues were segmented with a Dice score value of more than 0.90, while 89% of the cerebral areas had a DSC of more than 0.80.

IMPLICATIONS FOR PATIENT CARE: The developed software tool increases the amount of information extracted from standard, whole-body PET/CT datasets and provides means to perform system-level investigations into human physiology.

REFERENCES

1. Cannon WB. The wisdom of the body. *Am J Med Sci*. 1932;184:864.

2. Goodman L. Regulation and control in physiological systems: 1960-1980. *Ann Biomed Eng*. 1980;8:281-290.

3. Billman GE. Homeostasis: The Underappreciated and Far Too Often Ignored Central Organizing Principle of Physiology. *Front Physiol*. 2020;11:200.

4. Lammertsma AA. Forward to the Past: The Case for Quantitative PET Imaging. *J Nucl Med*. 2017;58:1019-1024.

5. Cherry SR, Badawi RD, Karp JS, Moses WW, Price P, Jones T. Total-body imaging: Transforming the role of positron emission tomography. *Sci Transl Med*. 2017;9:eaaf6169.

6. Karp JS, Viswanath V, Geagan MJ, et al. PennPET Explorer: Design and Preliminary Performance of a Whole-Body Imager. *J Nucl Med*. 2020;61:136-143.

7. Spencer BA, Berg E, Schmall JP, et al. Performance evaluation of the uEXPLORER total-body PET/CT scanner based on NEMA NU 2-2018 with additional tests to characterize PET scanners with a long axial field of view. *J Nucl Med*. 2021;62:861-870.

8. Prenosil GA, Sari H, Fürstner M, et al. Performance Characteristics of the Biograph Vision Quadra PET/CT System with a Long Axial Field of View Using the NEMA NU 2-2018 Standard. *J Nucl Med*. 2022;63:476-484.

9. Zhu L, Finkelstein D, Gao C, et al. Multi-organ Mapping of Cancer Risk. *Cell*. 2016;166:1132-1146.e7.

10. Penet M-F, Winnard PT Jr, Jacobs MA, Bhujwalla ZM. Understanding cancer-induced cachexia: imaging the flame and its fuel. *Curr Opin Support Palliat Care*. 2011;5:327-333.

11. Argilés JM, Busquets S, Stemmler B, López-Soriano FJ. Cancer cachexia: understanding the molecular

396 basis. *Nat Rev Cancer*. 2014;14:754-762.

397 12. Priest C, Tontonoz P. Inter-organ cross-talk in metabolic syndrome. *Nat Metab*. 2019;1:1177-1188.

398 13. Gupta A, Madhavan MV, Sehgal K, et al. Extrapulmonary manifestations of COVID-19. *Nat Med*.
399 2020;26:1017-1032.

400 14. Wu A. A Chat with Andrew on MLOps: From Model-Centric to Data-Centric AI. 2021.

401 15. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for
402 deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203-211.

403 16. Motamedi M, Sakharnykh N, Kaldewey T. A Data-Centric Approach for Training Deep Neural
404 Networks with Less Data. *arXiv [csAI]*. October 2021.

405 17. Badawi RD, Shi H, Hu P, et al. First Human Imaging Studies with the EXPLORER Total-Body PET
406 Scanner. *J Nucl Med*. 2019;60:299-303.

407 18. Weston AD, Korfiatis P, Kline TL, et al. Automated Abdominal Segmentation of CT Scans for Body
408 Composition Analysis Using Deep Learning. *Radiology*. 2019;290:669-679.

409 19. Traub-Weidinger T, Muzik O, Sundar LKS, et al. Utility of Absolute Quantification in Non-lesional
410 Extratemporal Lobe Epilepsy Using FDG PET/MR Imaging. *Front Neurol*. 2020;11:54.

411 20. Sundar LKS, Muzik O, Rischka L, Hahn A. Promise of Fully Integrated PET/MRI: Noninvasive Clinical
412 Quantification of Cerebral Glucose Metabolism. *Journal of Nuclear*. 2020.

413 21. Hammers A, Allom R, Koepp MJ, et al. Three-dimensional maximum probability atlas of the human
414 brain, with particular reference to the temporal lobe. *Hum Brain Mapp*. 2003;19:224-247.

415 22. Ashburner J. SPM: a history. *Neuroimage*. 2012;62:791-800.

416 23. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an

algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004;23:903-921.

24. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26:297-302.

25. Yeghiazaryan V, Voiculescu I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J Med Imaging (Bellingham)*. 2018;5:015006.

26. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv [csLG]*. December 2019.

27. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The Design of SimpleITK. *Front Neuroinform*. 2013;7:45.

28. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30:1323-1341.

29. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278:563-577.

30. NCCIH strategic plan FY 2021–2025. NCCIH. https://www.nccih.nih.gov/about/nccih-strategic-plan-2021-2025.

31. Hacker M, Hicks RJ, Beyer T. Applied Systems Biology—embracing molecular imaging for systemic medicine. *Eur J Nucl Med Mol Imaging*. 2020;47:2721-2725.

32. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. ; 2009:248-255.

ETHICS DECLARATION

All the data utilized in this study were acquired in accordance with the Declaration of Helsinki. Written informed consent was obtained from all the subjects prior to the examinations.

Dataset	Acquisition system	Institutional Review Board	Reference number
34 ^{18}F-FDG PET/MR brain datasets	Siemens Biograph mMR, Siemens Healthineers	Medical University of Vienna	EK1960/2014
30 low-dose healthy control CT datasets	uEXPLORER, United Imaging Healthcare	The University of California at Davis	I1341792-18
20 low-dose mixed pathological Siemens CT datasets	Siemens Biograph mCT TruePoint TrueV, Siemens Healthineers	Medical University of Vienna	EK1649/2016
Three Lymphoma datasets	Philips Gemini GXL16, Philips Medical Systems	Protection des Personnes Sud-Est III, Hôpital HOTEL-DIEU, Place de l'Hôpital	Etude REMARC Réf: 2009 - 006B; Eudract N°: 2008-008202-52.
Three mesothelioma datasets	Siemens Biograph mCT TruePoint TrueV, Siemens Healthineers	Medical University of Vienna	EK1649/2016

FIGURES

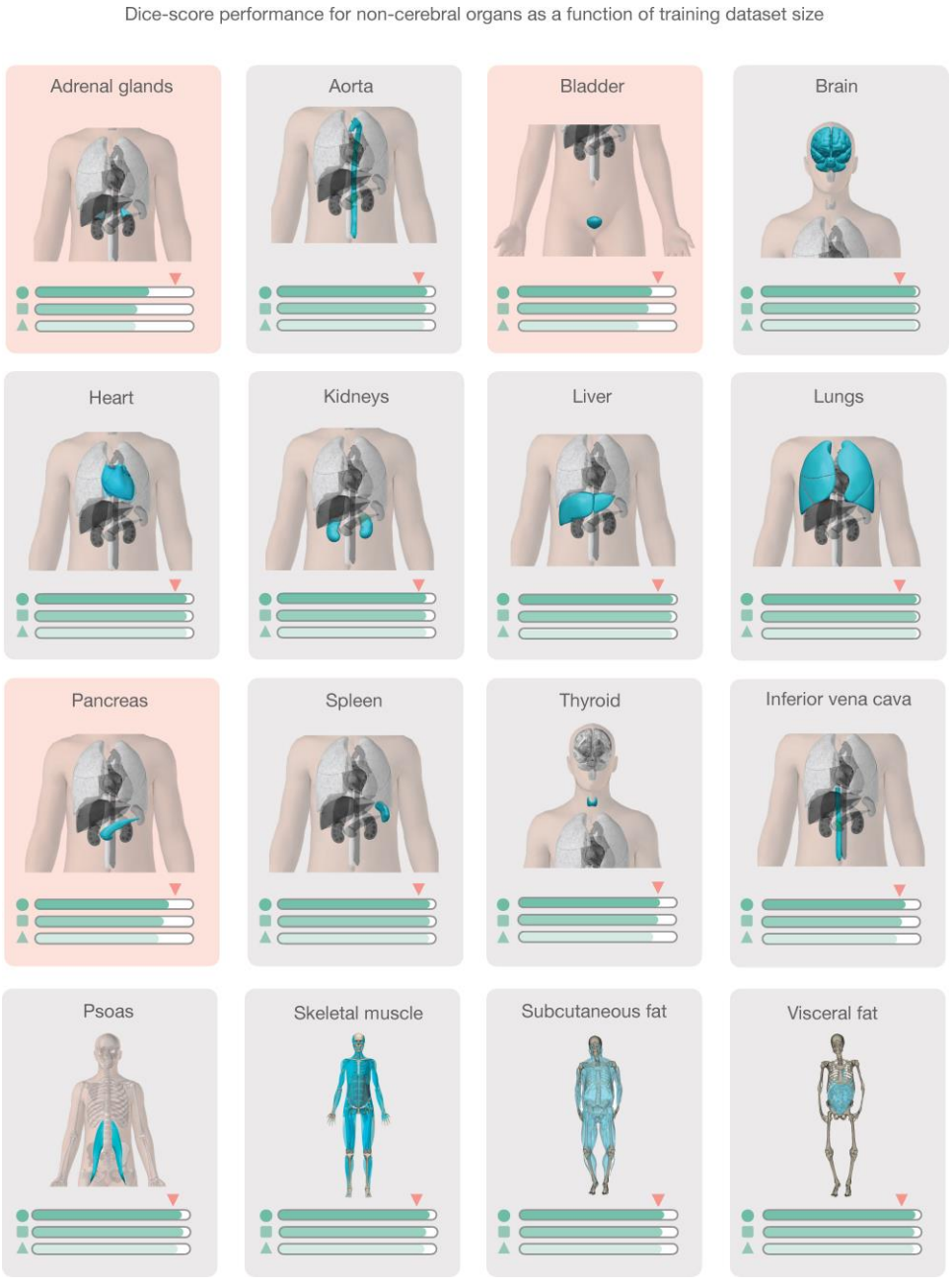


Figure 1. Median DSCs of abdominal organs (10 test datasets) were obtained from models based on three separate training subsets: D40 (circle), D20 (square), D10 (triangle). The inverted triangle (pink) indicates the 0.90 mark. Red background highlights organs characterised by low median DSCs (<0.90) and high standard deviation (SD, see **Supplemental Fig. 4).**

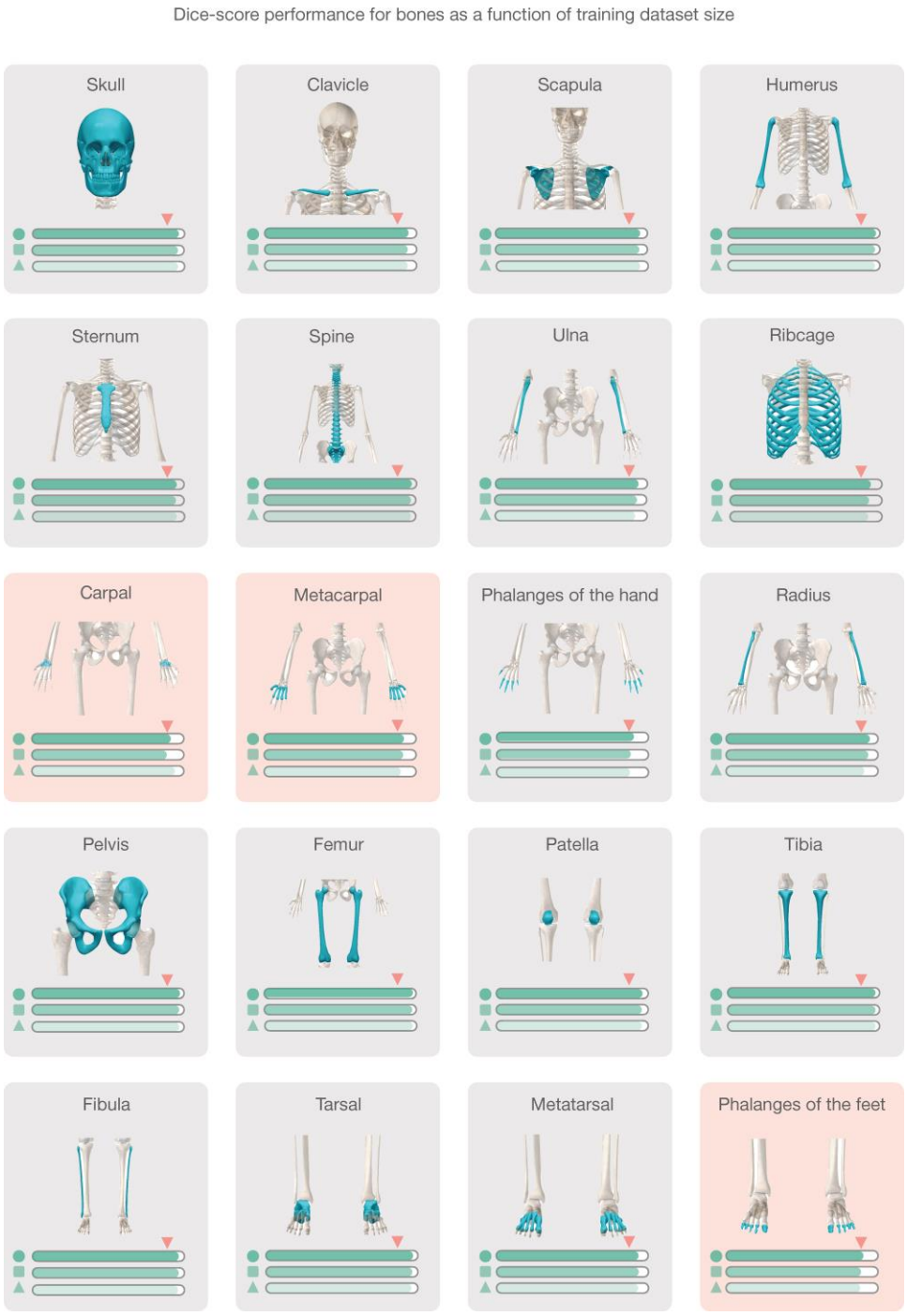


Figure 2. Median DSCs of bone structures (10 test datasets) as obtained from models based on three separate training subsets: D40 (circle), D20 (square), and D10 (triangle). The inverted triangular marker indicates the 0.90 mark. Red background highlights bones characterised by low median DSCs (<0.90).

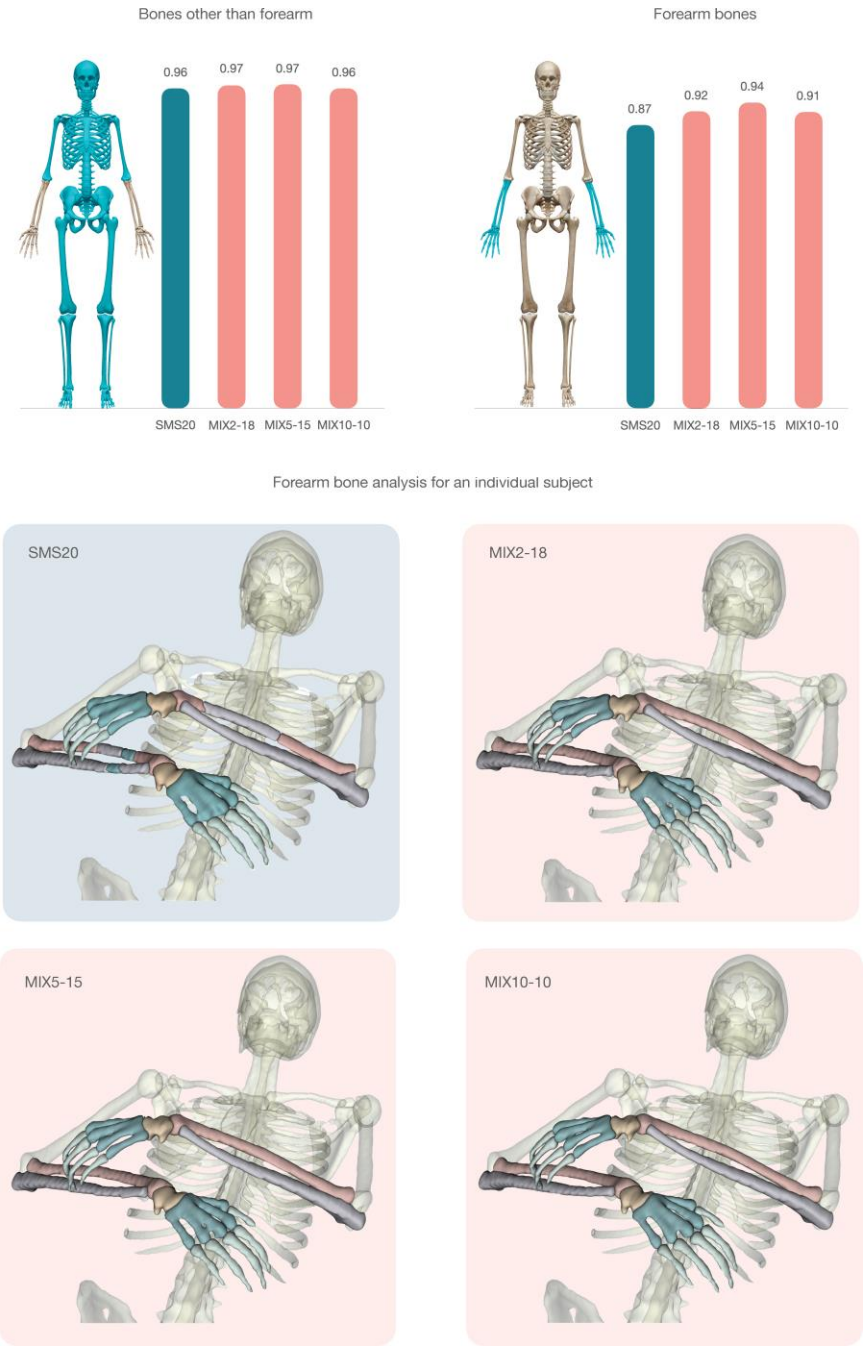


Figure 3. (A) Bar graph demonstrating a similar performance of different models for bone segmentation other than forearm bones. Green bar depicts the homogenous training dataset (SMS-20), whereas mixed training datasets (MIX2-18, MIX5-15, MIX10-10) are represented by red bars. (B) Bar graph showing segmentation performance of forearm bones. A significant performance improvement is seen in the mixed training datasets (red bars) compared to the homogeneous training dataset (green bar). (C) Forearm bone analysis of an individual subject. The images demonstrate that the forearm bones are incorrectly segmented in the case of the SMS20 (green background) model, whereas all mixed models accurately segmented the forearm bones (red background).

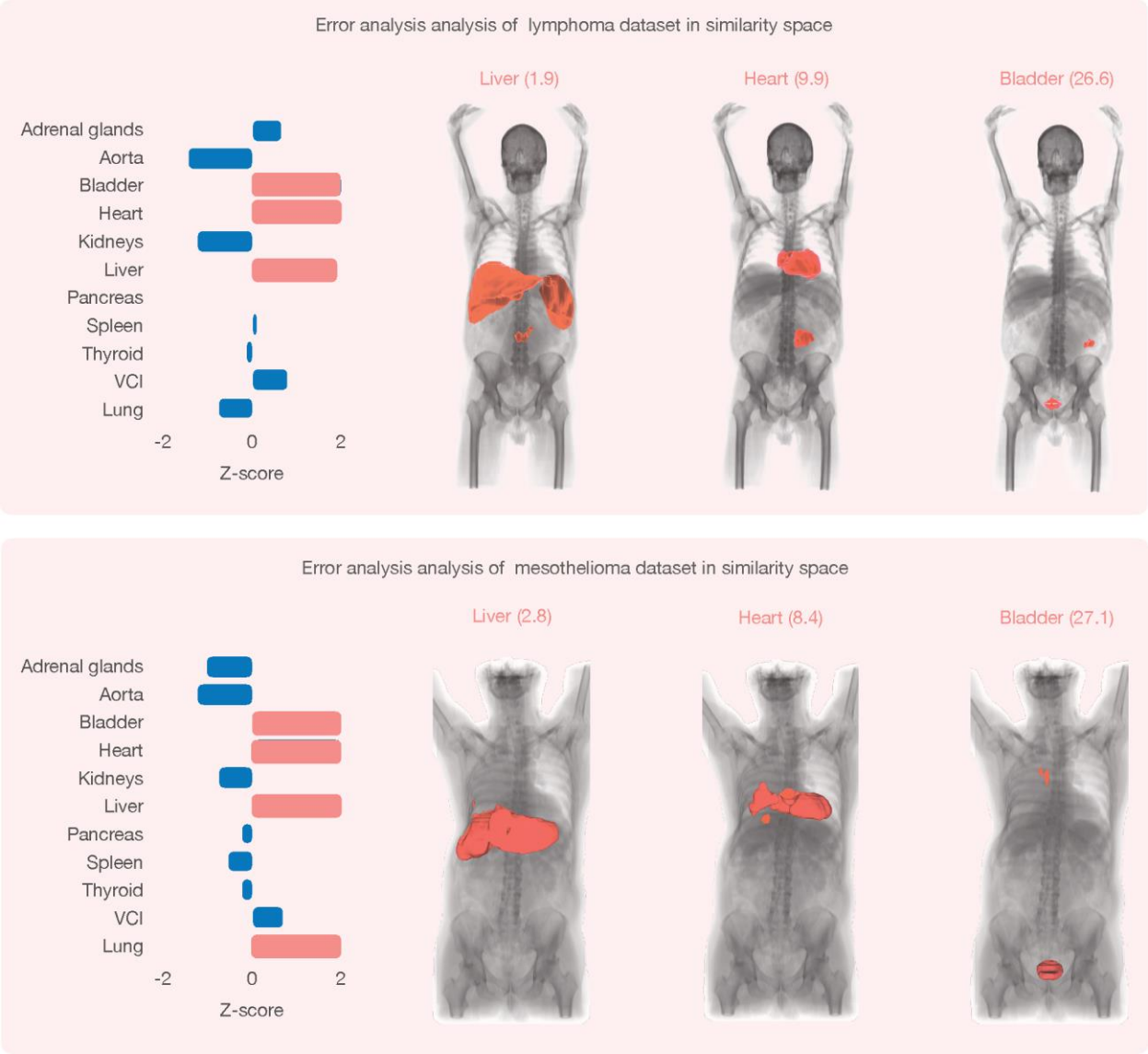
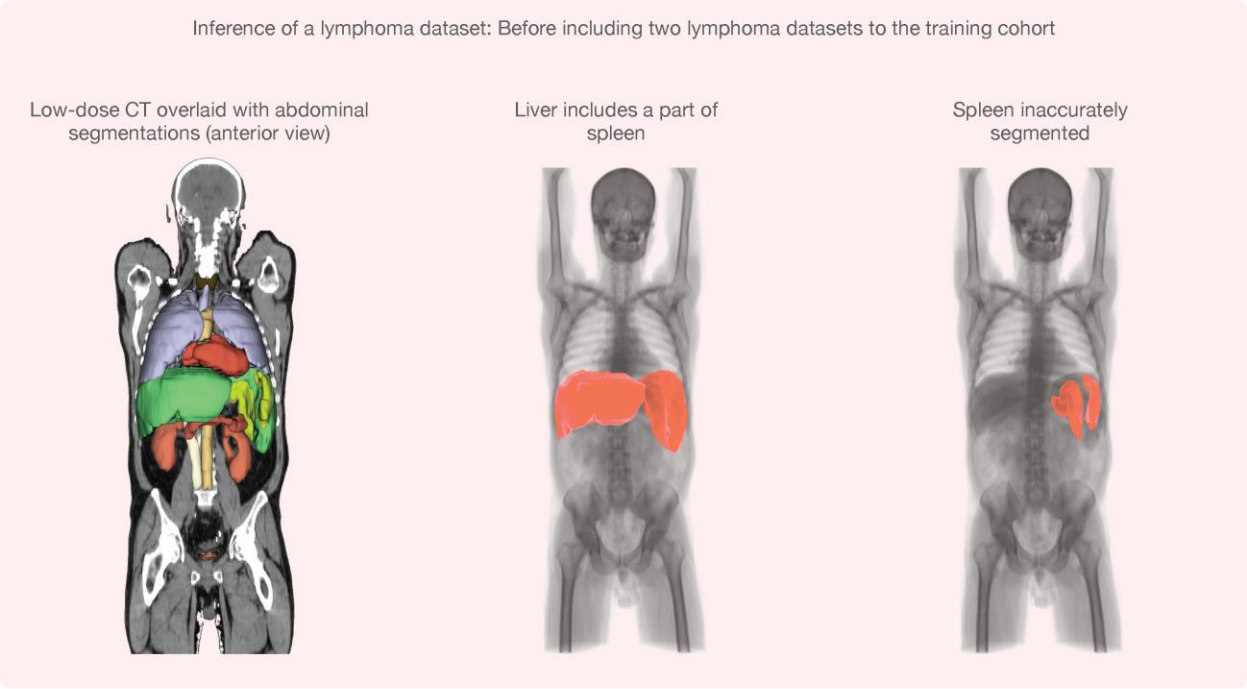
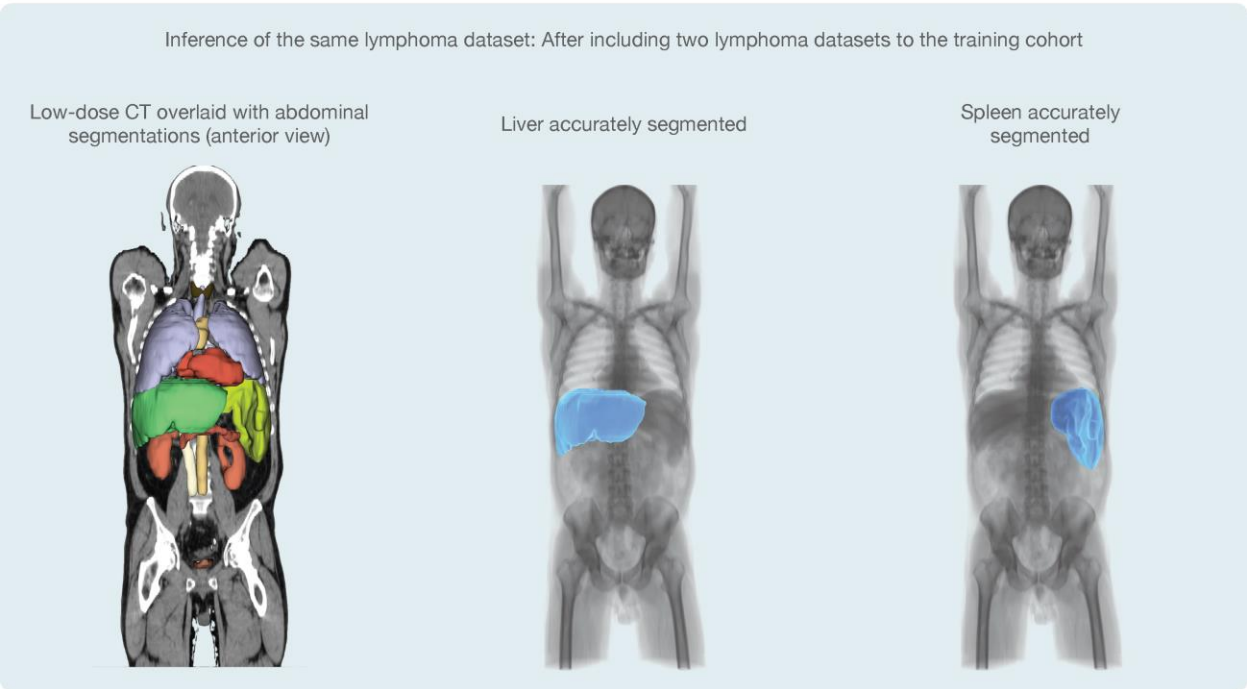


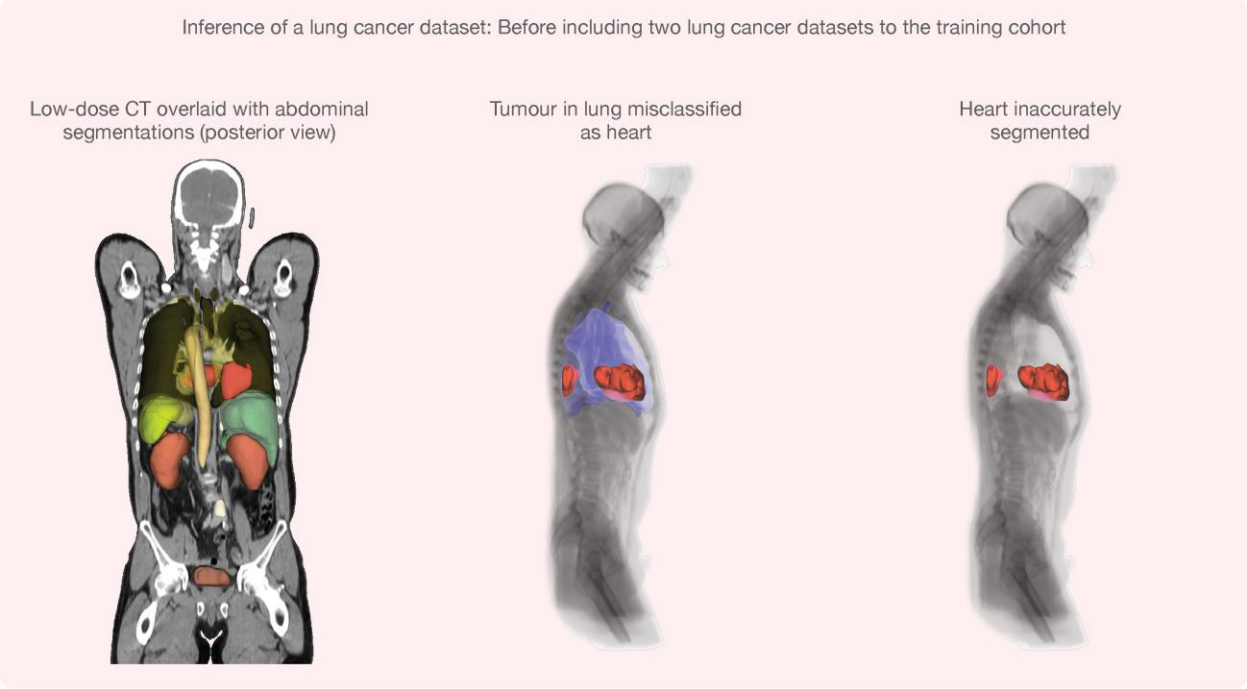
Figure 4 (A) Error analysis in similarity space for a representative lymphoma patient. Horizontal bars depict distance in similarity space, with blue bars characterizing organs with a z-value of <1.5. The figure shows z-scores >1.5 for the liver, kidneys, and bladder (red bars). Corresponding organ segmentations are displayed to the right for the liver (z = 1.9) and the heart (z = 9.9), indicating suboptimal segmentation results that require manual correction. (B) Error analysis in similarity space for a representative mesothelioma patient with z-scores >1.5 for the liver, heart, bladder, and lung. Incorrect organ segmentations are shown to the right for the liver (z = 2.8) and the heart (z = 8.4).

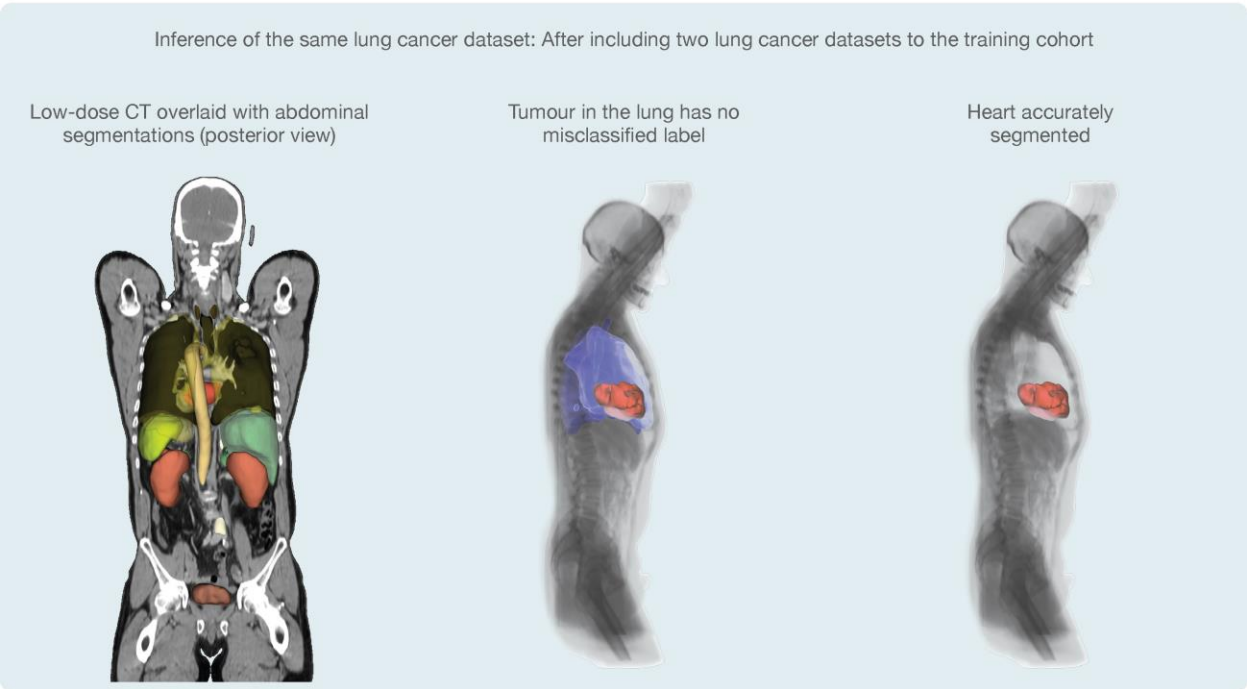




469

470 **Figure 5** (A) Organ segmentation of a hold-out lymphoma test dataset using a training dataset that did
471 not include splenomegaly cases. (B) Organ segmentation of the same patient following inclusion of 2
472 (different) lymphoma datasets and model retraining using the expanded training dataset. It can be seen
473 that the updated model is able to recognise the new image pattern resulting in a correct segmentation of
474 both the liver and the spleen.
475

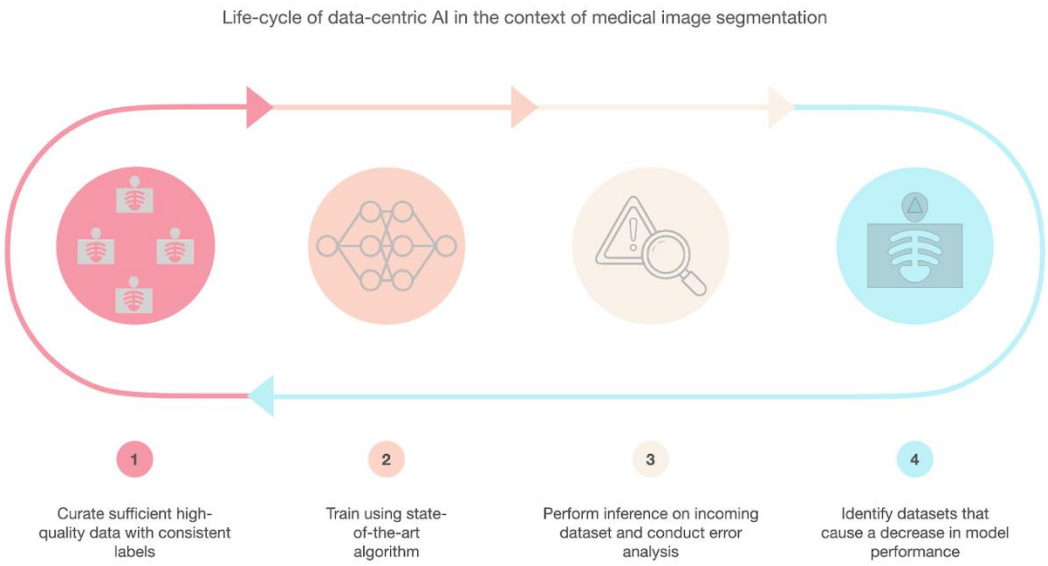




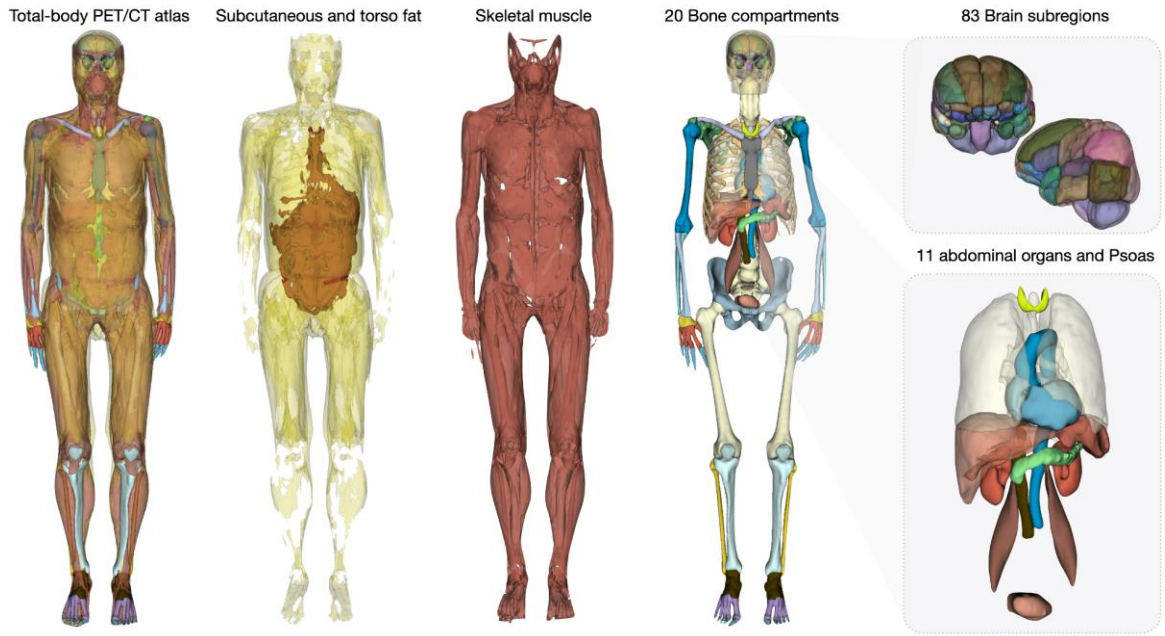
476

477 **Figure 6** (A) Organ segmentation of a hold-out mesothelioma test dataset using a training dataset that did
478 not include mesothelioma cases. (B) Organ segmentation of the same patient following inclusion of 2
479 (different) mesothelioma datasets and model retraining using the expanded training dataset. The updated
480 model recognized the new image pattern resulting in a correct segmentation of the heart.
481

SUPPLEMENTAL FIGURES



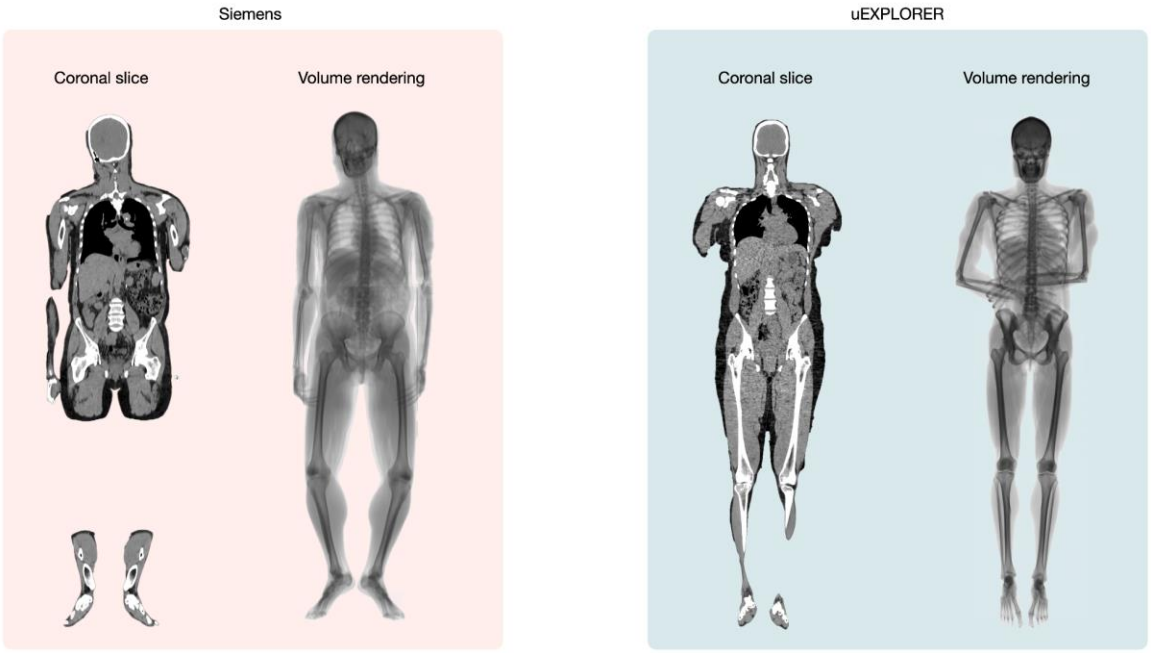
Supplemental Figure 1. Life-cycle of data-centric AI approach. The main goal is to identify the datasets that cause a reduction in the model's performance. Once these datasets are identified, they are added to the original training dataset, and the network is retrained.



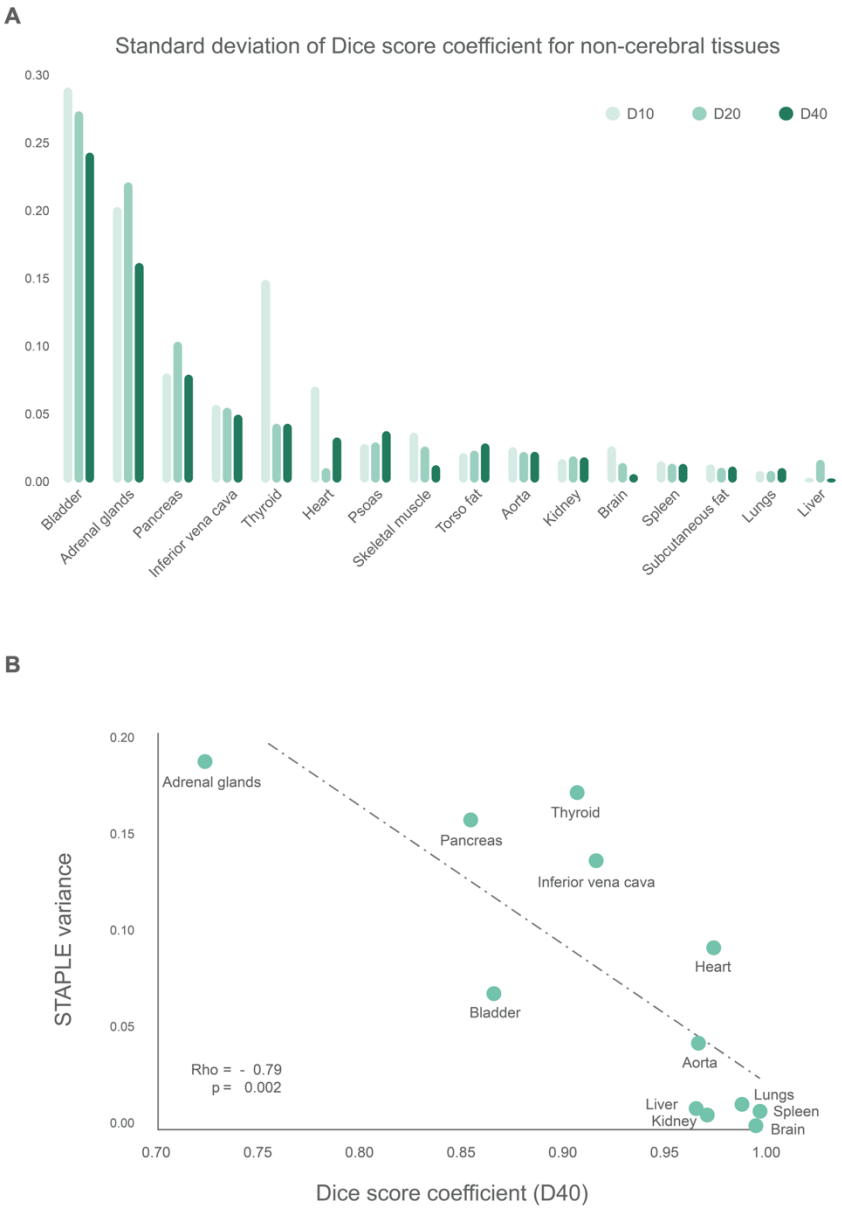
490

491 **Supplemental Figure 2**. ^{18}F-FDG total-body PET/CT tissue-map consisting of 120 unique tissues - both
492 cerebral and non-cerebral structures.

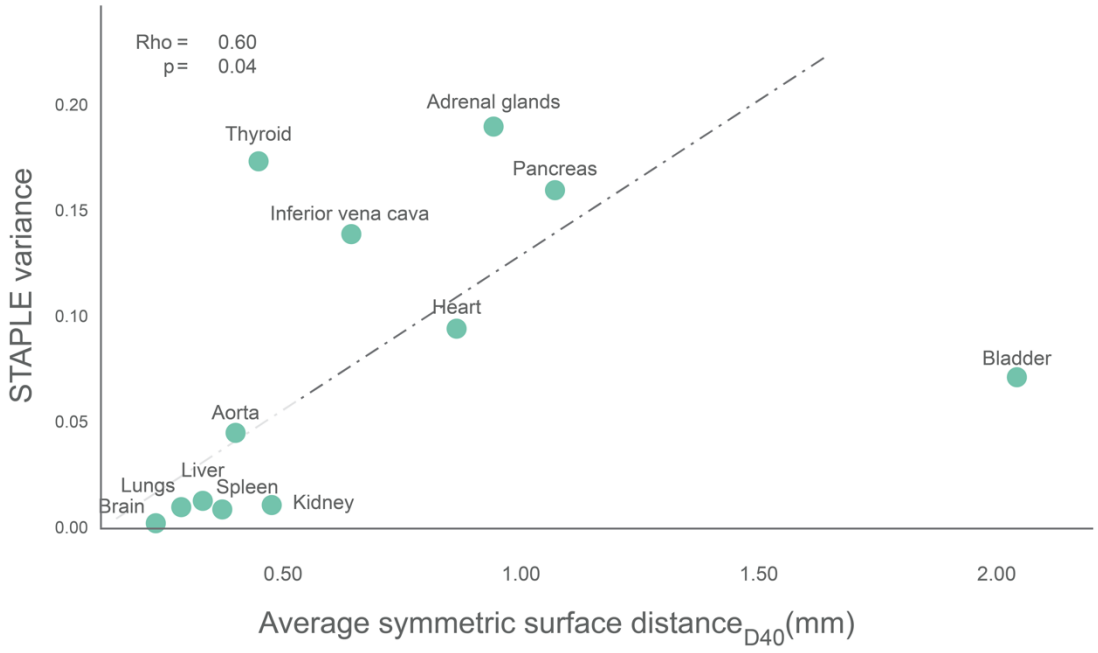
493



Supplemental Figure 3. An example image of a low-dose CT from the Siemens (left) and uEXPLORER (right) system. The two images differ in their noise characteristics and the hand-positions of the subjects. Subjects scanned using the Siemens system position their hands side-by-side (left), while subjects scanned using the uEXPLORER system have their hands crossed across the chest (right).



Supplemental Figure 4. **(A)** Standard deviation (SD) values of Dice Score Coefficients (DSCs) were determined for various abdominal organs derived using 10 (D10), 20 (D20) and 40 (D40) training datasets. The graph indicates that SD values are similar for most organs except for the bladder, adrenal gland, thyroid and pancreas. **(B)** The correlation plot shows a highly significant negative correlation between the DSCs obtained from the D40 training dataset and the STAPLE variance. The organs with the lowest DSC values are also those with the highest SD (adrenal gland, pancreas and thyroid), indicating that segmentation of these three structures is problematic even for human experts.

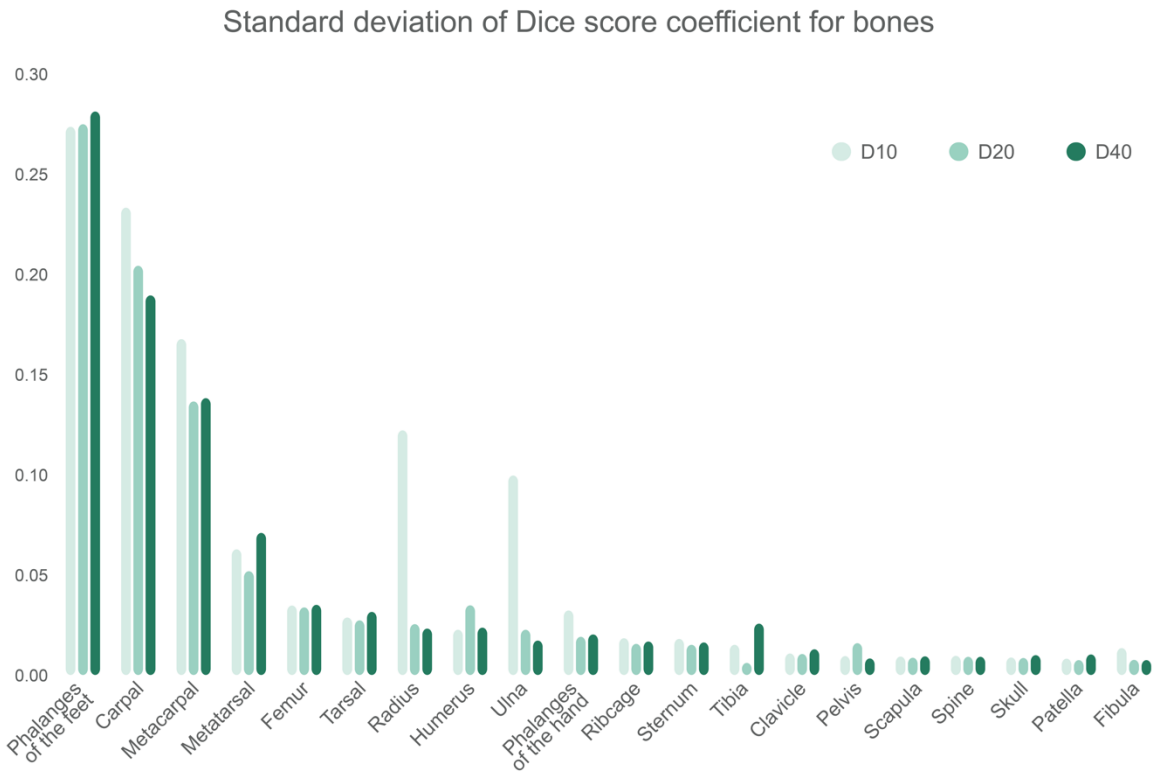


509

510 **Supplemental Figure 5**. The correlation plot shows a significant correlation between the average
511 symmetric surface distance (ASSD) and the STAPLE variance, indicating that the segmentation of the
512 bladder, pancreas and the adrenal gland is more challenging than for other organs. It can be seen that the
513 average misalignment between contours is about twice (>1mm) that of other organs, confirming the
514 conclusion derived using the DSC metric.

515

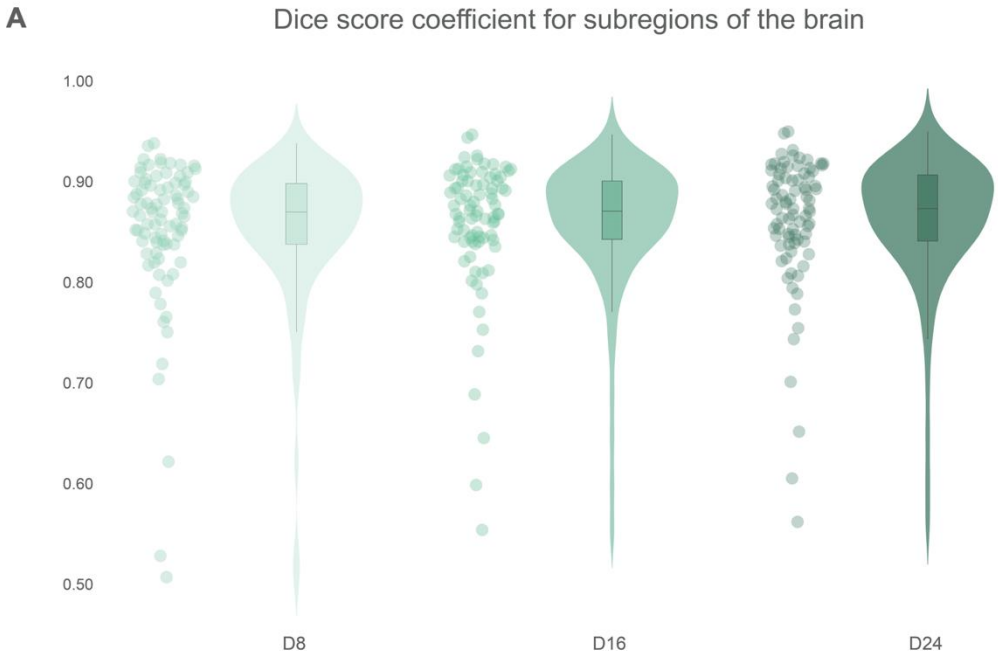
516

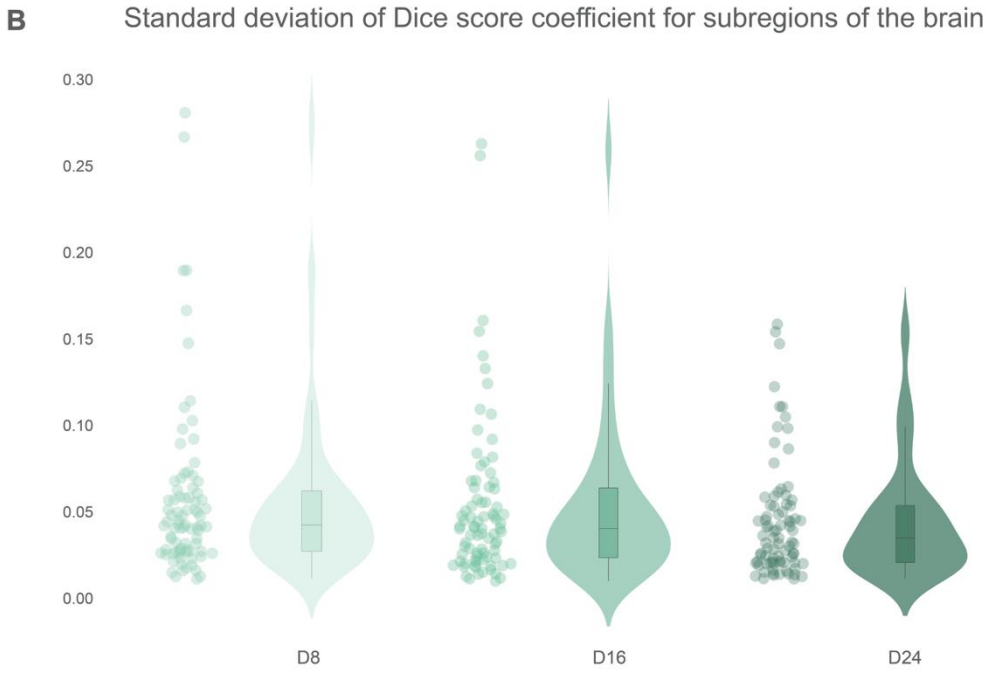


517

518 **Supplemental Figure 6**. Standard deviation (SD) values of Dice Score Coefficients (DSCs), determined for
519 various bone structures derived using 10 (D10), 20 (D20) and 40 (D40) training data sets. The graph
520 indicates that the SD of the three data sets is similar for most organs, except for the radius and ulna. The
521 variability in segmentation accuracy of these two bone structures is high in case of a small training data.

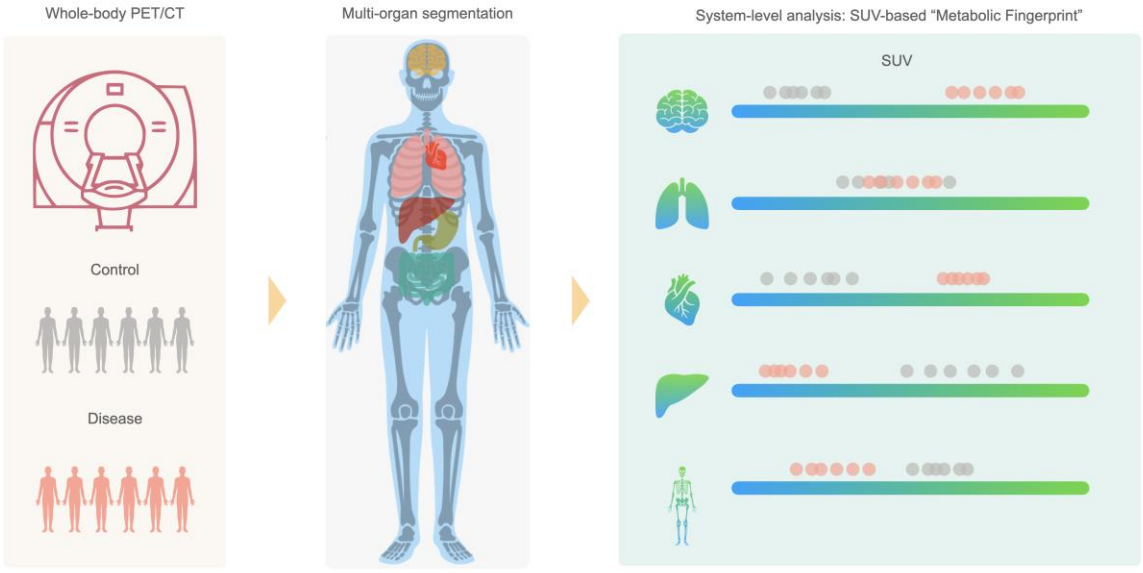
522





523

524 **Supplemental Figure 7**. Mean (A) and SD (B) values of the DSCs determined for various brain regions
525 (denoted by dots) derived using 8 (D8), 16 (D16) and 24 (D24) training datasets. One can appreciate that
526 increase in the training data set size has only a minor effect on segmentation performance, causing a small
527 decrease in the DSC variance.



528

529 **Supplemental Figure 8**. Concept diagram showing the extraction of a "metabolic fingerprint" from PET/CT
530 data obtained from a control and disease population. The developed automated multi-organ
531 segmentation pipeline is used to extract organ SUV values from both groups (control: grey, disease:
532 orange), allowing the study of systems-level alterations in the pattern of organ metabolic activity as a
533 consequence of the disease process.
534

SUPPLEMENTAL TABLES

Supplemental Table 1. List of cerebral (highlighted in green) and non-cerebral tissues (highlighted in orange) along with their corresponding label-index in the total-body digital FDG PET/CT tissue-map.

Label-index	Tissues
1	R-Hippocampus
2	L-Hippocampus
3	R-Amygdala
4	L-Amygdala
5	R-Anterior-temporal-lobe-medial-part
6	L-Anterior-temporal-lobe-medial-part
7	R-Anterior-temporal-lobe-lateral-part
8	L-Anterior-temporal-lobe-lateral-part
9	R-Parahippocampal-and-ambient-gyri
10	L-Parahippocampal-and-ambient-gyri
11	R-Superior-temporal-gyrus-posterior-part
12	L-Superior-temporal-gyrus-posterior-part
13	R-Middle-and-inferior-temporal-gyrus
14	L-Middle-and-inferior-temporal-gyrus
15	R-Fusiform-gyrus
16	L-Fusiform-gyrus
17	R-Cerebellum
18	L-Cerebellum
19	Brainstem
20	L-Insula
21	R-Insula
22	L-Lateral-remainder-of-occipital-lobe
23	R-Lateral-remainder-of-occipital-lobe
24	L-Cingulate-gyrus-gyrus-cinguli-anterior-part
25	R-Cingulate-gyrus-gyrus-cinguli-anterior-part
26	L-Cingulate-gyrus-gyrus-cinguli-posterior-part
27	R-Cingulate-gyrus-gyrus-cinguli-posterior-part

28	L-Middle-frontal-gyrus
29	R-Middle-frontal-gyrus
30	L-Posterior-temporal-lobe
31	R-Posterior-temporal-lobe
32	L-Inferiolateral-remainder-of-parietal-lobe
33	R-Inferiolateral-remainder-of-parietal-lobe
34	L-Caudate-nucleus
35	R-Caudate-nucleus
36	L-Nucleus-accumbens
37	R-Nucleus-accumbens
38	L-Putamen
39	R-Putamen
40	L-Thalamus
41	R-Thalamus
42	L-Pallidum
43	R-Pallidum
44	Corpus-callosum
45	R-Lateral-ventricle-excluding-temporal-horn
46	L-Lateral-ventricle-excluding-temporal-horn
47	R-Lateral-ventricle-temporal-horn
48	L-Lateral-ventricle-temporal-horn
49	Third-ventricle
50	L-Precentral-gyrus
51	R-Precentral-gyrus
52	L-Straight-gyrus
53	R-Straight-gyrus
54	L-Anterior-orbital-gyrus
55	R-Anterior-orbital-gyrus
56	L-Inferior-frontal-gyrus
57	R-Inferior-frontal-gyrus
58	L-Superior-frontal-gyrus
59	R-Superior-frontal-gyrus
60	L-Postcentral-gyrus

61	R-Postcentral-gyrus
62	L-Superior-parietal-gyrus
63	R-Superior-parietal-gyrus
64	L-Lingual-gyrus
65	R-Lingual-gyrus
66	L-Cuneus
67	R-Cuneus
68	L-Medial-orbital-gyrus
69	R-Medial-orbital-gyrus
70	L-Lateral-orbital-gyrus
71	R-Lateral-orbital-gyrus
72	L-Posterior-orbital-gyrus
73	R-Posterior-orbital-gyrus
74	L-Substantia-nigra
75	R-Substantia-nigra
76	L-Subgenual-frontal-cortex
77	R-Subgenual-frontal-cortex
78	L-Subcallosal-area
79	R-Subcallosal-area
80	L-Pre-subgenual-frontal-cortex
81	R-Pre-subgenual-frontal-cortex
82	L-Superior-temporal-gyrus-anterior-part
83	R-Superior-temporal-gyrus-anterior-part
84	Adrenal-glands
85	Aorta
86	Bladder
87	Brain
88	Heart
89	Kidneys
90	Liver
91	Pancreas
92	Spleen

93	Thyroid
94	Inferior vena cava
95	Lung
96	Carpal
97	Clavicle
98	Femur
99	Fibula
100	Humerus
101	Metacarpal
102	Metatarsal
103	Patella
104	Pelvis
105	Phalanges of the hand
106	Radius
107	Ribcage
108	Scapula
109	Skull
110	Spine
111	Sternum
112	Tarsal
113	Tibia
114	Phalanges of the feet
115	Ulna
116	Skeletal-muscle
117	Subcutaneous-fat
118	Torso-fat
119	Psoas
120	Entire Skeleton

538

GRAPHICAL ABSTRACT